Computing and Artificial Intelligence 2024, 2(1), 441.
https://doi.org/10.59400/cai.v2i1.441

Article

Identifying voices using convolution neural network models AlexNet and ResNet

Abdulaziz Alhowaish Luluh, Muniasamy Anandhavalli*

College of Computer Science, King Khalid University, Abha 62529, Kingdom of Saudi Arabia
*** Corresponding author:** Anandhavalli Muniasamy, anandhavalli.dr@gmail.com

CITATION

Luluh AA, Anandhavalli M. Identifying voices using convolution neural network models AlexNet and ResNet. Computing and Artificial Intelligence. 2024; 2(1): 441. https://doi.org/10.59400/cai.v2i1.441

ARTICLE INFO

Received: 2 January 2024
Accepted: 30 January 2024
Available online: 12 February 2024

Abstract: Deep learning (DL) techniques which implement deep neural networks became popular due to the increase of high-performance computing facilities. DL achieves higher power and flexibility due to its ability to process many features when it deals with unstructured data. DL algorithm passes the data through several layers; each layer is capable of extracting features progressively and passes it to the next layer. Initial layers extract low-level features, and succeeding layers combine features to form a complete representation. This research attempts to utilize DL techniques for identifying sounds. The development in DL models has extensively covered classification and verification of objects through images. However, there have not been any notable findings concerning identification and verification of the voice of an individual from different other individuals using DL techniques. Hence, the proposed research aims to develop DL techniques capable of isolating the voice of an individual from a group of other sounds and classify them based on the use of convolutional neural networks models AlexNet and ResNet, that are used in voice identification. We achieved the classification accuracy of ResNet and AlexNet model for the problem of voice identification is 97.2039 % and 65.95% respectively, in which ResNet model achieves the best result.

Keywords: deep learning (DL); voice identification; convolutional neural network (CNN); AlexNet; ResNet

1. Introduction

Due to the continual advancement of artificial intelligence (AI) technology, deep learning (DL) has become increasingly important in the last decade. DL is an artificial intelligence (AI) structure that impersonates the activities of the human mind in handling information for use in identifying objects, perceiving speech, interpreting dialects, and making decisions based on the data collected [1]. Deep learning refers to a machine-learning algorithm that can locate objects, identify speech, translate, and interpret languages, and make decisions without the need for human supervision. So, deep learning has the advantage that the software builds the feature set independently and without supervision. While deep learning techniques have been successfully implemented in recognizing characters and objects, its development in identifying voices and interpreting them is still under process. Deep learning achieves this by employing a multi-layered calculation design known as neural networks. The design of the human brain influences the neural organizing plan. In the same way that humans think carefully to discern designs and order various types of data, neural networks may be trained to do the same thing with data. The individual layers of neural organizations can likewise be considered as such a channel that works from gross to unpretentious, improving the probability of identifying and yielding a right outcome [2]. While the AI's performance in voice recognition has been widely praised, it is also necessary to

address the AI's issues in distinguishing the voices of individuals. As voice-based AI systems are now included in personal security and essential surveillance technologies, the requirement for distinguishing voices is higher than ever. The current research used deep learning techniques to attempt to recognize voices in this regard. It focuses primarily on the usage of convolutional neural networks, which provide non-linearity toward an audio-based set, making it easier to operate. Moreover, according to Deng and Liu [3], the relevance of identifying and interpreting voices has grown in recent years. The increasing importance is due to an increase in demand for voice-controlled AI systems in the technology.

The human brain works in the same way with each point involving humans accessing new data, the brain attempts to contrast it to other known items. A similar idea is additionally utilized by deep neural networks where the systems use the data collected to be compared to already existing information [4]. Deep learning is notable for its relevance in image recognition. However, another critical utilization of the innovation is in speech recognition utilized in digital assistances such as Amazon's Alexa or messaging through voice recognition [5]. The benefit of deep learning in voice identification comes from the adaptability and flexibility that comes from deep neural networks that lately become more accessible. Meanwhile, A critical challenge in identifying the right words involves the variability of the voice produced for the same word given in different accents such as Hello vs. Hellooo. When issued with an audible sentence the voice recognition starts by altering the voice waves using Fast Fourier Transformation and concatenating structures from adjacent windows in order to come up with a spectrogram [6]. The main aim is to take down the dimensionality of the univariate voice data enabling prediction of the letters and words that are coming next. Computer oriented processing and the recognition of human voices is referred to as speech recognition. It can be used to validate user credentials in certain systems as well as give instructions to digital assistants such as Siri, Cortana or Google Assistant. These systems work through storage of human voices which then train the voice identification system to listen in for vocabulary and patterns within the speech produced.

Another vital test in voice identification is the issue of redundancy; to decipher progressively, the model should foresee words accurately without the entire sentence. A portion of the deep learning models like bi-directional repetitive neural networks advantage exceptionally from utilizing the entire sentence because of the additional unique [7]. The arrangement in diminishing inactivity is to remember restricted setting for the model construction by permitting the neural organization to approach a short measure of data after a particular word [8]. At last, deep learning is still genuinely in its outset, however, is rapidly moving toward a cutting-edge capacity in voice identification. Studies indicate that there is still a great deal of opportunity to get better in model designing to diminish inertness and increment precision [9].

Though object verification and classification of images have benefited greatly from the widespread use of deep learning, the identification and verification of individual voices remains conspicuously lacking. Many DL developments to far have gone toward translating and understanding voice instructions, which has left a gap in the capacity to identify and authenticate individual voices within a group. In order to close this gap, this article uses deep learning methods for voice identification, namely

convolutional neural networks models like AlexNet and ResNet. Voice recognition for individuals involves a machine consuming voices, speeches, utterances, and phrases from an individual so that each time the individual speaks, the speech is converted to voice waves and compared to previously stored datasets in order to recognize the specific person. It is important to note that the identification of voices through deep learning improves with time and the measure of data fed into the system. The architecture of deep learning, such as AlexNet and ResNet models and the network of the adversarial generative and recurrent neural networks for memory and long-term as obstetric non-supervised models. Neural Networks sets the background of the AlexNet and ResNet as models discriminatory moderated, presented and developed it. The application of this structure on a large scale to many applications such as automatic speech, natural language processing, and the establishment of voice and detection of objects in three-dimensional where it is made proven that it produces results developed for different tasks.

In summary, this paper makes the following contributions:

- Apply convolutional neural network models and evaluate the performance in identifying and verifying the voice of a particular speaker.
- Explore ResNet and AlexNet models in identifying the voice.
- Evaluate the model's results and compare them with other deep learning or machine learning models.

2. Literature review

Khalil et al. [10] is presenting different deep learning techniques for speech emotion recognition (SER) as it is known that the recognition of emotion is a very challenging part of the human computer interaction (HCI). Extraction of the emotion from the voice signal is being done using many techniques, which is classification techniques and a well-established analysis of the speech. All the traditional techniques now have been replaced by the deep learning technique for speech emotion recognition. A convolution neural network has been used for the classification feed-forward architecture (**Figure 1**).

In **Figure 1**, the basic layer-wise architecture of the convolution neural network is presented.

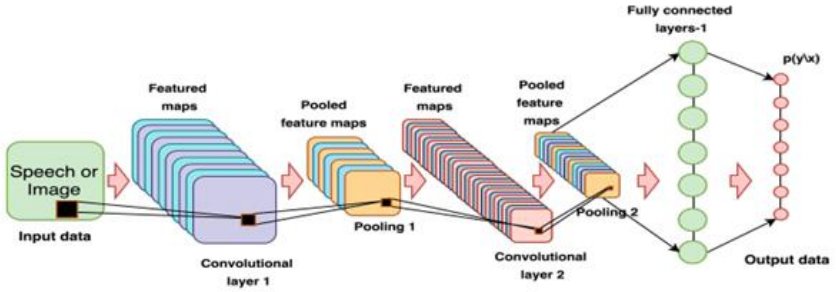

Figure 1. Architecture of CNN layer-wise.

Recognition of the pattern is done with the help of a convolution neural network, and classification of the data in a better form is provided by the CNN. In each layer small size of a neuron are present in these types of networks, the layer is of architecture

which is designed model which is used for the processing receptive field which is a type of input data [10].

On the other hand, Zhang et al. [11] point to the use of convolutional neural network as an effective method to be involved with deep learning techniques for recognizing sounds. They believe that since convolutional neural networks have multiple hidden layers, the users could set more parameters for refining the sounds that pass through the AI learning process. As in agreement, Deng and Liu [3] says that by representing the audio input using the spectrogram image, it is possible to identify the minute the changes in the hertz and decibels of a sound. In this regard, Deng and Liu [3] carried on an analysis that involved isolation of specific sounds from multiple sounds using the convolutional neural network. By identifying the spectrogram plots, the network was successful in isolating the specific sound. However, Lemley et al. [12] argue that though the approach would yield expected results in isolating environment sounds, it would not be able to distinguish between the words spoken by a person. The reason is that, regarding human speech, the spectrogram plots are hard to identify, as they look similar to one another. Even so, Parcollet et al. [13] stated in their study that convolutional neural network when used with deep learning techniques would enable an AI to improve its classification parameters and identify sounds of individuals, given that a specific speech input has already been made.

In Zhang et al. [14] they focus on the details of the Adam optimization algorithm, which is utilized to solve the objective function iteratively. As a result, by constantly modifying the first order and second-order momentum with an average of gradients over iteration, the Adam algorithm has great strength in resisting noise pollution.

In ShabanAl-Ani [15], fingerprint recognition is one of the most famous biometrics. Due to its excellence and consistency over time, fingerprints have been used for identification for more than a century. In this study, discrete cosine transform (DCT) is applied to the entire image. DCT communicates a limited succession of items as far as an aggregate of cosine capacities swaying at various frequencies. This gives DCT transactions, which are universal fingerprint recognition policy features based on separate pocket conversion technology, suggested. This algorithm relies on the matching process through the link in the images, and then uses DCT technology to extract important features. This study faces many problems such as the low quality of the fingerprint input image.

According to Krizhevsky et al. [16], AlexNet famously won the 2012 ImageNet LSVRC-2012 competition by a large margin (15.3% vs. 26.2% (second place) error rates). The network had a very similar architecture to LeNet but was deeper, with more filters per layer, and with stacked convolutional layers.

In their study on sound classification using AI, Andrew [17] figured out that computers could read a sound better if it were to be given an image representation of those sounds. As such, Andrew [17] suggests carrying on a visual inspection of the audio files and then differentiating them based on the difference in their graphs. Malik et al. [18] argue that using the image representation of sound waves could offer inaccurate results in classifying sounds. In their study carried out urban sound classification, the study realized that, for certain sounds, the visual representation remains almost identical to one another. In such cases, it becomes hard to differentiate one sound from another.

In another study, Pons et al. [19] conducted an experiment on isolating environmental sounds using deep learning techniques. The study made use of a neural network but with a single hidden layer. The neural network was placed as a middle layer to filter the sounds based on given parameters. The experiment was successful in isolating the sounds in the environment. However, as it was stated by Bunrit et al. [20], using a single hidden layer limits the number of parameters that the user could set in recognizing the sounds. While it may be successful in differentiating sounds with notable level of frequency, it cannot isolate mild sounds that the environment may also contain.

According to the author Heaven [21], the categorization of speaker verification and speaker identification can be done in the speaker recognition approach. A spectrogram image is being used for the training of the CNN method used in the identification and verification of the voice and is best than the other system used. 95.83% was the average classification result of the method used for the testing. In the case of the MFCC method, the average was 91.26% and 49.77% for the raw signal as an input to the CNN model. So, when the shot voice is used as an input, then the method is very efficient.

3. Methodology

This research aims to use convolutional neural networks models, AlexNet and ResNet, to identify the voices of the individual in an environment and classify voice dataset to achieve the best accuracy of the voice identification. In this section we explained the methodology of our work. **Figure 2** shows the proposed methodology.

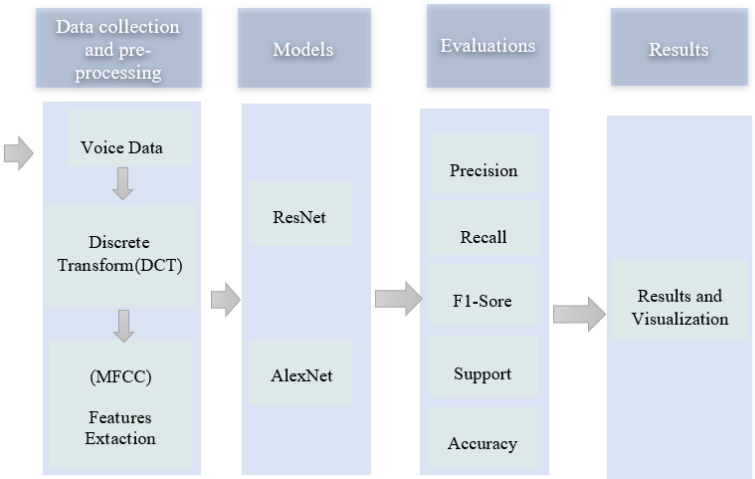

Figure 2. Proposed methodology structure.

Qualitative research methodology uses non-numerical data for understanding the underlying reasons behind selected subject [22]. Meanwhile, quantitative research methodology encourages the use of numerical data, it can help in identifying patterns and making predictions [23]. Our research requires the use of numerical data and recognition of patterns in the changes with regard to voices and their classification. Hence, we used quantitative research methodology. The data required for carrying on the analysis is an existing voice dataset from OpenSLR website. We used MFCC to extract features from voices, and then we applied the ResNet and AlexNet models for

the training and testing process. The varying levels of accuracy in identifying the specific set would enable the current research to understand the level of applicability of CNN in identifying voices.

The following are the steps to implement the proposed methodology of voice identification of the speaker:

Step 1: Collection of voice datasets contains 2937 audio files.

Step 2: Feature extraction using MFCC.

Step 3: Divide the dataset into train and test sets as a ratio of 9:1.

Step 4: Set labels IDs for recognition.

Step 5: Training the ResNet and AlexNet models for about 50 epochs for better prediction.

Step 6: Save the accurate trained models to test voices.

3.1. Dataset description

We used raw human voice dataset from LibriSpeech ASR corpus [24]. LibriSpeech is a corpus of approximately 1000 h and 16 kHz of English speeches, prepared by Vassil Panayotov with the assistance of Daniel Povey.

;ID	SEX	SUBSET	MINUTES	NAME
14	F	train-clean-360	25.03	Kristin LeMoine
16	F	train-clean-360	25.11	Alys AtteWater
17	M	train-clean-360	25.04	Gord Mackenzie
19	F	train-clean-100	25.19	Kara Shallenberg
. ...				

Figure 3. Dataset description.

The data is derived from read audiobooks from the LibriVox project and has been carefully segmented and aligned. The used dataset is “dev-clean.tar.gz”, that contains 2937 audio of people and is considered as voice classification of 40 classes, and the samples are shown in **Figure 3**. The samples of shape of audio signals are shown in **Figure 4**.

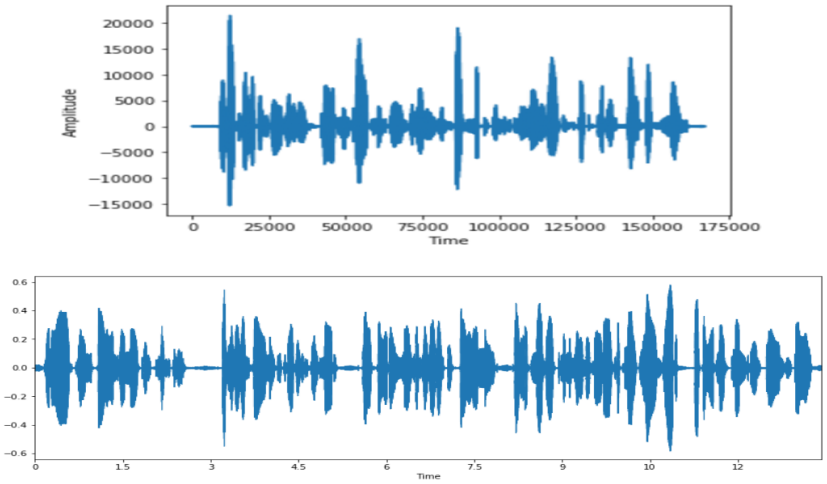

Figure 4. Sample of shape of original audio signals.

3.2. MFCC (Mel-Frequency Cepstral Coefficients)

The main process in voice identification is speech pre-processing. We used MFCC (Mel-Frequency Cepstral Coefficients) as a tool for extracting voice dynamics functions as acoustic features (such as MFCCs) as inputs and speaker IDs as target variable (**Figure 5**).

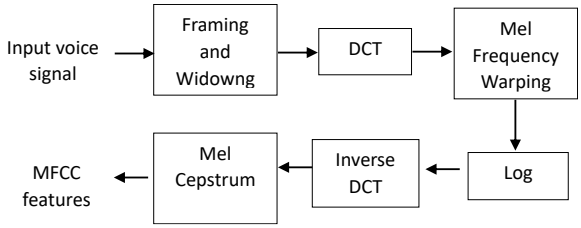

Figure 5. MFCC process flow diagram.

Voice signals in the time domain change very quickly, so we convert speech signals from the time domain to the frequency domain, then the corresponding spectrum can be clearly defined. We separate the signals into frames and calls the window function to increase the continuity of voice signals in the frame. Discrete cosine transform (DCT) is being used for quantitative evaluation of spectral energy data into data units that can be analyzed by MFCC [25]. The famous audio processing library "librosa 0.7.2" has been used to compute the MFCC features of audios (**Figure 6**). The sampling rate is set to 16,000 and the number of MFCCs is set to 24.

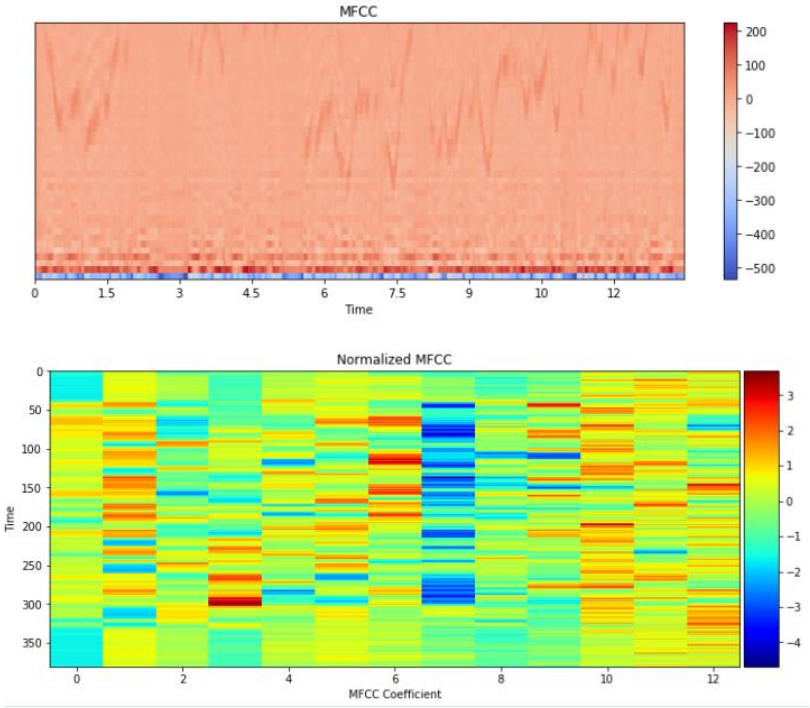

Figure 6. Sample of MFCC extracted features of audio.

We can observe that more no. of filters in low frequency region, and lesser no. of filters in high frequency region. Finally, these extracted features that we used in training and voice identification model using ResNet and AlexNet.

3.3. CNN models for classification

The ResNet (residual neural network) consists of many CNN layers, batch norm layers, max pooling layers, and ReLu layers. The core idea of ResNet is introducing a so-called "identity shortcut connection" that skips one or more layers, so it's very good at classification tasks and its block model is shown in **Figure 7**.

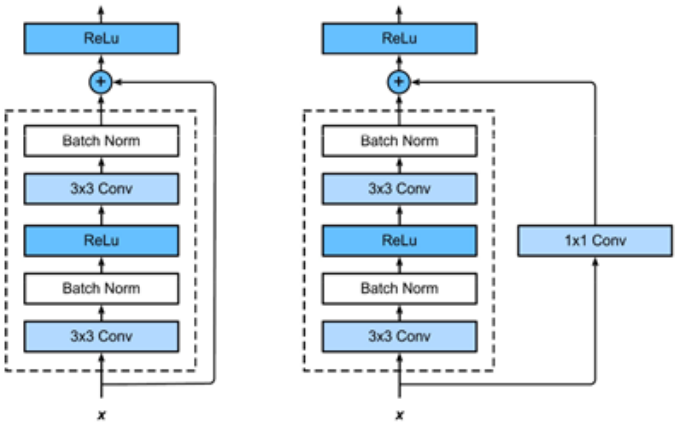

Figure 7. ResNet block model.

Figure 7 shows the ResNet consists of many CNN layers, batch norm layers, max pooling layers, and ReLu layers. The core idea of ResNet is introducing a so-called "identity shortcut connection" that skips one or more layers, so it's very good at classification tasks. ResNet model consists of 11 ResNet blocks, and a ResNet block consists of 3 CNN layers and batch norm and max pooling layers have been applied for this research.

AlexNet (**Figure 8**) is a convolutional neural network which has had a large impact on the field of machine learning, specifically in the application of deep learning to machine vision.

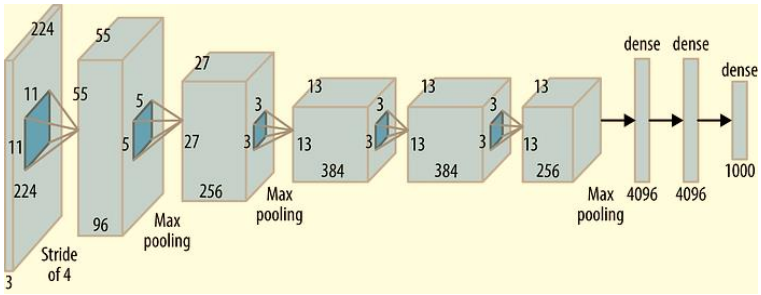

Figure 8. Architecture of AlexNet.

AlexNet consists of 11 × 11, 5 × 5, 3 × 3, convolutions, max pooling, dropout, data augmentation, ReLU activations, SGD with momentum. It attached ReLU activations after every convolutional and fully connected layer. AlexNet was trained simultaneously on two Nvidia Geforce GTX 580 GPUs which is the reason why their network is split into two pipelines.

3.4. Evaluations

Classification evaluations metrics (**Figure 9**) includes accuracy, precision, and

recall. Precision is measured the positive patterns that are correctly predicted from the total predicted patterns in a positive class. Recall is used to measure the fraction of positive patterns that are correctly classified, and the accuracy metric measures the ratio of correct predictions over the total number of instances evaluated. F-score derived from two measures which are precision and recall, this metric represents the harmonic mean between recall and precision values [26].

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

Figure 9. Description of metrics.

True positive (TP) means the patient has the condition and the test was positive. True negative (TN) indicates that the patient is healthy with a negative test result. A false negative (FN) indicates when negative samples are incorrectly expected to be positive. False positive (FP) refers to when positive patient samples are incorrectly anticipated to be negative.

4. Results and discussion

In this section, a description of the experimental results and performance evaluation, which shows the effectiveness of models that extract results of experiment and comparison between results with graphs and tables has been discussed. The proposed model was built with python deep learning framework: "PyTorch 1.2", the code runs in google colab.

Training parameters:

The dataset is split into training and test sets at a ratio of 9:1, which is the most effective split ratio for our learning process. In this research we used the loss function which is the Adam optimization. Adam optimization algorithm is an extension to stochastic gradient descent that has recently seen broader adoption for deep learning applications. **Table 1** shows Hyperparameters values that control the learning process.

Table 1. Hyperparameters for data training.

No.	Hyperparameter	Value
1	Learning rate	0.0005
2	Batch size	16
3	Num of epochs	50
4	Loss function	Adam

The performance of deep learning is affected by hyperparameters. Also, they are critical for the efficacy of the optimization and model-fitting process [27]. So, we used the appropriate values to train our CNN models that gave the best performance. The chosen hyperparameters are the Learning rate, the ratio into training and validation is

0.0005, batch size is 16, the size of a batch must be more than or equal to one and less than or equal to the number of samples in the training dataset. Also, the number of epochs is 50, the number of complete passes through the training dataset.

Figure 10 shows the training (loss, accuracy) for ResNet model with 50 iterations from 1 to 49. Epochs is the number of complete passes through the training dataset, each iteration show loss, accuracy and val-accuracy for make improvement in ResNet model to arrive the greatest accuracy and lowest loss in our ResNet model with loss 0.0669% and higher accuracy 97.2039% and val-accuracy is 88.970% respectively.

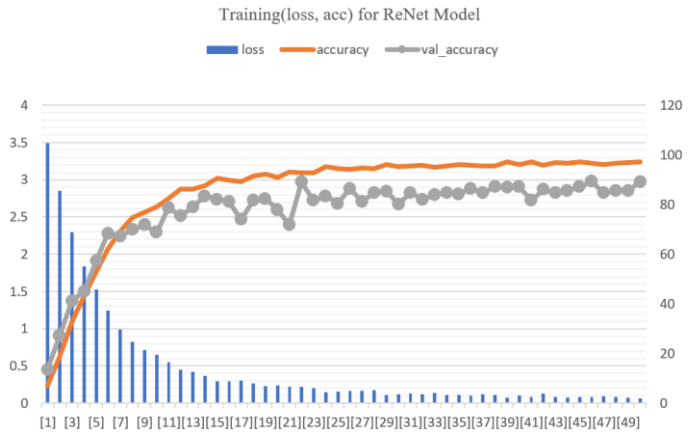

Figure 10. Training (loss, acc) for ResNet model.

Figure 11 shows the information about the accuracy of the ResNet model in each epoch (50 epochs) or iteration along with the loss is in the graph, by increased epoch number lowered loss value. The *X*-axis of the graph is the epochs, and the *Y*-axis is the result percentage. The minimum loss value is 0.0259, best training accuracy is 98.64%, and best validation accuracy is 87.86%. Also, we show the relation between training loss and accuracy for data using the ResNet model. Training loss value in data reduced and the validation accuracy is still lower than training accuracy then this model is excellent for our system, this means our model is fitted.

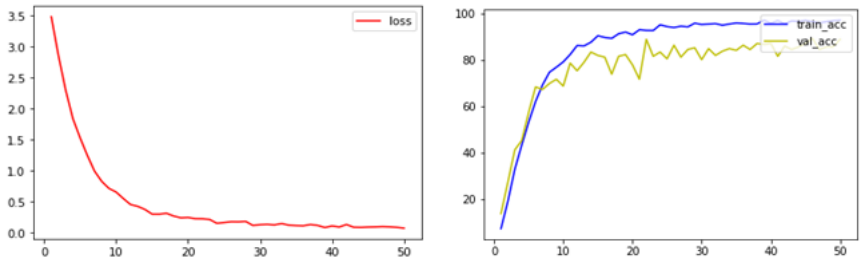

Figure 11. Line plots for (Training loss, Train-accuracy and Val-accuracy) of the ResNet model.

Classification report of the ResNet model is given in **Figure 12**. Report shows the results of 40 number of data voices from dataset to identify this voice belonged to whom. So, the test classification report that belongs to the speaker with an ID 84, ResNet model shows 90% accuracy, precision value 83%, recall value 100% and F1-score is 91%.

>>>>>>>>>>>> Classification report >>>>>>>

	precision	recall	f1-score	support
3576	1.00	0.50	0.67	6
777	0.85	1.00	0.92	11
2035	1.00	0.70	0.82	10
3170	1.00	0.80	0.89	5
422	1.00	1.00	1.00	3
3000	1.00	1.00	1.00	6
174	0.80	1.00	0.89	8
2086	1.00	1.00	1.00	1
1993	0.75	0.75	0.75	4
2078	1.00	1.00	1.00	3
5895	1.00	1.00	1.00	6
652	0.89	1.00	0.94	8
8842	1.00	0.67	0.80	3
2277	0.78	0.88	0.82	8
84	0.83	1.00	0.91	5
1462	1.00	0.88	0.93	8
2428	0.83	1.00	0.91	10
6313	1.00	1.00	1.00	11
5694	0.92	0.92	0.92	13
3536	1.00	1.00	1.00	4
5338	1.00	0.67	0.80	6
3853	1.00	1.00	1.00	7
6241	1.00	0.67	0.80	9
1919	0.80	1.00	0.89	4
1272	1.00	1.00	1.00	4
251	0.89	0.73	0.80	11
2412	1.00	1.00	1.00	3
8297	0.83	1.00	0.91	10
7976	0.75	0.75	0.75	4
2902	0.67	1.00	0.80	2
2803	1.00	1.00	1.00	8
6319	1.00	0.75	0.86	8
3752	1.00	0.90	0.95	10
3081	0.92	1.00	0.96	11
6345	0.71	0.92	0.80	13
6295	1.00	1.00	1.00	3
7850	0.83	0.83	0.83	6
1673	0.50	0.67	0.57	3
1988	1.00	1.00	1.00	8
5536	1.00	1.00	1.00	6
accuracy			0.90	269
macro avg	0.91	0.90	0.90	269
weighted avg	0.91	0.90	0.90	269

Figure 12. Classification report ResNet model.

Figure 13 shows training (loss, acc) for AlexNet model with 50 iteration, each iteration show loss, accuracy and val-accuracy for make improvement in AlexNet model to arrive the greatest accuracy and lowest loss with loss 1.064% and accuracy 65.95% respectively. Compared to ResNet model classification results AlexNet model does not show the good accuracy in this research.

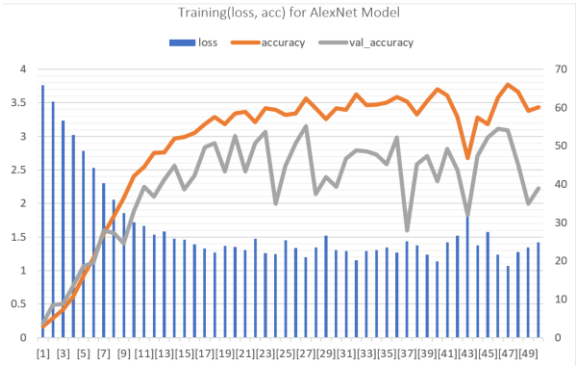

Figure 13. Training (loss, acc) for AlexNet model.

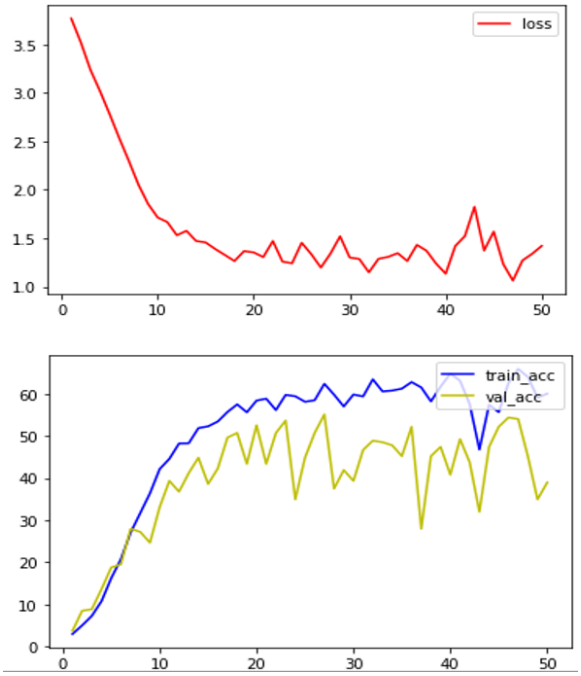

Figure 14. Line plots for (Training loss, Train-accuracy and Val-accuracy) of the AlexNet model.

Figure 14 shows the information about the accuracy of the AlexNet model in each epoch (50 epochs) or iteration along with the loss is in the graph. The *x*-axis of the graph is the epochs and the *y*-axis is the result percentage. The minimum loss value is 1.064%, best training accuracy is 65.95%, and best validation accuracy is 54.04%. The validation accuracy is still lower than training accuracy. So, from achieved results of the two models, ResNet model gives better accuracy than AlexNet model.

Classification report of the AlexNet model is given in **Figure 15**. Report shows the results of data voices from dataset to identify this voice belonged to whom. So, the test classification report that belongs to the speaker with an ID 1988, AlexNet model shows 39% accuracy.

Figure 16 shows the average of classification metrics of the ResNet and AlexNet Models. Based on the results, the ResNet model produces a higher average by precision 91.3% and recall 89.9 % and F1-score 89.7 %. We found that the results of the AlexNet model with classification metrics values of precision 42.5% and recall 41.2 % and F1-score 37.1% respectively.

```
predict_audio(alexnet_model,"./dataset/dataset/1988/1988-147956-0000.flac")

calc_metrics(alexnet_model)

This audio is belonged to  1988
>>>>>>>>>>>> Classification report >>>>>>>
              precision    recall  f1-score   support

        3576       0.00      0.00      0.00         6
         777       0.50      0.36      0.42        11
        2035       0.88      0.70      0.78        10
        3170       0.50      0.60      0.55         5
         422       0.00      0.00      0.00         3
        3000       0.12      0.33      0.18         6
         174       0.23      0.38      0.29         8
        2086       1.00      1.00      1.00         1
        1993       0.29      0.50      0.36         4
        2078       0.00      0.00      0.00         3
        5895       0.86      1.00      0.92         6
         652       1.00      0.50      0.67         8
        8842       0.00      0.00      0.00         3
        2277       0.29      0.62      0.40         8
          84       0.42      1.00      0.59         5
        1462       0.38      0.38      0.38         8
        2428       0.00      0.00      0.00        10
        6313       0.92      1.00      0.96        11
        5694       0.44      0.31      0.36        13
        3536       0.67      0.50      0.57         4
        5338       0.67      0.33      0.44         6
        3853       0.50      0.14      0.22         7
        6241       0.20      0.11      0.14         9
        1919       0.14      0.50      0.22         4
        1272       0.22      0.50      0.31         4
         251       0.15      0.55      0.24        11
        2412       0.50      0.33      0.40         3
        8297       0.00      0.00      0.00        10
        7976       0.50      1.00      0.67         4
        2902       0.33      0.50      0.40         2
        2803       0.00      0.00      0.00         8
        6319       1.00      0.25      0.40         8
        3752       0.50      0.10      0.17        10
        3081       1.00      0.36      0.53        11
        6345       0.64      0.69      0.67        13
        6295       0.23      1.00      0.38         3
        7850       0.43      0.50      0.46         6
        1673       1.00      0.33      0.50         3
        1988       0.00      0.00      0.00         8
        5536       0.50      0.17      0.25         6

    accuracy                           0.39       269
   macro avg       0.43      0.41      0.37       269
weighted avg       0.44      0.39      0.37       269
```

Figure 15. Classification report AlexNet model.

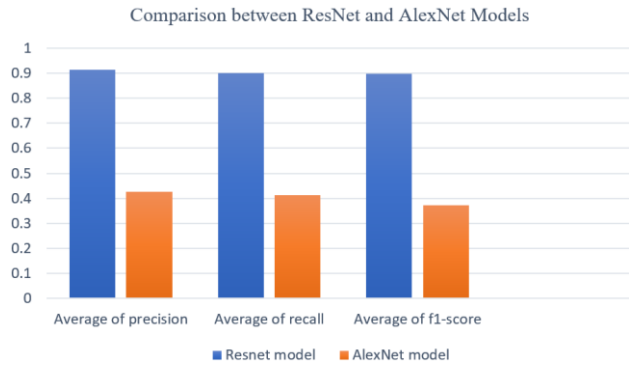

Figure 16. Average of classification metrics of the ResNet and AlexNet models.

Table 2 presents a comparison of the proposed work and other studies related to the identification of voices using classification algorithms.

Based the comparison results, the ResNet model's results can assist with the problem of identifying the voices as ResNet model achieved the results of precision value 91.3%, recall value 89.9%, and F1-score is 89% and AlexNet model achieves

the results of precision 42.5 %, recall 41.3 %, and F1-score 37.1%, respectively.

Table 2. Comparison between the proposed research and related work.

Study	Classification models	Accuracy
Proposed work	**ResNet**	**97.2039%**
	AlexNet	**65.95%**
Mamyrbayev O et al. [28]	Support vector method	83%
	Robust scaler method	90%
Bunrit S et al. [20]	CNN model	95.83%
	MFCC model	91.26%
Khalil R et al. [10]	Deep recurrent neural network	92.3%
Pons J et al. [19]	CNN model	88.95%

5. Conclusion

This study offers deep learning-based based voice recognition models. Convolutional neural networks have been used for identifying and verifying the speakers by their voices, and we tested and identified the voice of a particular speaker. In the initial step of data preparation, we employed the MFCC algorithm in the voice preprocessing process. ResNet model and AlexNet, models of convolutional neural network have been applied for voice identification. Classification accuracy of ResNet and AlexNet model for the problem of voice identification is 97.2039 % and 65.95% respectively. In the future work, we planned to incorporate voice verification technology into a multi-level and hybrid authentication approach to improve authentication reliability.

Author contributions: Conceptualization, AAL and MA; methodology, AAL; software, AAL; validation, AAL, and MA; formal analysis, MA; investigation, AAL; resources, AAL; data curation, AAL; writing—original draft preparation, AAL; writing—review and editing, MA; visualization, AAL; supervision, MA; project administration, MA. All authors have read and agreed to the published version of the manuscript.

Conflict of interest: The authors declare no conflict of interest.

References

1. Bae HS, Lee HJ, Lee SG. Voice recognition based on adaptive MFCC and deep learning. In 2016 IEEE 11th Conference on Industrial Electronics and Applications (ICIEA). 2016. pp. 1542-1546. doi: 10.1109/ICIEA.2016.7603830
2. Zhou R, Liu F, Gravelle CW. Deep Learning for Modulation Recognition: A Survey With a Demonstration. IEEE Access. 2020, 8: 67366-67376. doi: 10.1109/access.2020.2986330
3. Deng L, Liu Y. Deep Learning in Natural Language Processing. Springer Singapore, 2018. doi: 10.1007/978-981-10-5209-5
4. Deng L, Hinton G, Kingsbury B. New types of deep neural network learning for speech recognition and related applications: an overview. 2013 IEEE International Conference on Acoustics, Speech and Signal Processing. Published online May 2013. doi: 10.1109/icassp.2013.6639344
5. Terzopoulos G, Satratzemi M. Voice Assistants and Smart Speakers in Everyday Life and in Education. Informatics in Education. Published online September 16, 2020: 473-490. doi: 10.15388/infedu.2020.21
6. Nguyen G, Dlugolinsky S, Bobák M, et al. Machine Learning and Deep Learning frameworks and libraries for large-scale

data mining: a survey. Artificial Intelligence Review. 2019, 52(1): 77-124. doi: 10.1007/s10462-018-09679-z
7. Lee JG, Jun S, Cho YW, et al. Deep Learning in Medical Imaging: General Overview. Korean Journal of Radiology. 2017, 18(4): 570. doi: 10.3348/kjr.2017.18.4.570
8. Fayek HM, Lech M, Cavedon L. Evaluating deep learning architectures for Speech Emotion Recognition. Neural Networks. 2017, 92: 60-68. doi: 10.1016/j.neunet.2017.02.013
9. Bashar A. Survey on evolving deep learning neural network architectures. Journal of Artificial Intelligence and Capsule Networks. 2019, 2019(2): 73-82. doi: 10.36548/jaicn.2019.2.003
10. Khalil RA, Jones E, Babar MI, et al. Speech Emotion Recognition Using Deep Learning Techniques: A Review. IEEE Access. 2019, 7: 117327-117345. doi: 10.1109/access.2019.2936124
11. Zhang Y, Pezeshki M, Brakel P, et al. Towards End-to-End Speech Recognition with Deep Convolutional Neural Networks. Interspeech 2016. Published online September 8, 2016. doi: 10.21437/interspeech.2016-1446
12. Lemley J, Bazrafkan S, Corcoran P. Deep Learning for Consumer Devices and Services: Pushing the limits for machine learning, artificial intelligence, and computer vision. IEEE Consumer Electronics Magazine. 2017, 6(2): 48-56. doi: 10.1109/mce.2016.2640698
13. Parcollet T, Zhang Y, Morchid M, et al. Quaternion Convolutional Neural Networks for End-to-End Automatic Speech Recognition. Interspeech 2018. Published online September 2, 2018. doi: 10.21437/interspeech.2018-1898
14. Zhang Z. Improved Adam Optimizer for Deep Neural Networks. 2018 IEEE/ACM 26th International Symposium on Quality of Service (IWQoS). Published online June 2018. doi: 10.1109/iwqos.2018.8624183
15. ShabanAl-Ani M, M. Al-Aloosi W. Biometrics Fingerprint Recognition using Discrete Cosine Transform (DCT). International Journal of Computer Applications. 2013, 69(6): 44-48. doi: 10.5120/11849-7598
16. Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. Communications of the ACM. 2017, 60(6): 84-90. doi: 10.1145/3065386
17. Andrew. Nuts and bolts of building AI applications using Deep Learning. NIPS Keynote Talk. 2016.
18. Malik M, Adavanne S, Drossos K, et al. Stacked convolutional and recurrent neural networks for music emotion recognition. 2017.
19. Pons J, Slizovskaia O, Gong R, et al. Timbre analysis of music audio signals with convolutional neural networks. In: Proceedings of the 2017 25th European Signal Processing Conference (EUSIPCO) IEEE. pp. 2744-2748.
20. Bunrit S, Inkian T, Kerdprasop N, et al. Text-Independent Speaker Identification Using Deep Learning Model of Convolution Neural Network. International Journal of Machine Learning and Computing. 2019, 9(2): 143-148. doi: 10.18178/ijmlc.2019.9.2.778
21. Heaven D. Why deep-learning AIs are so easy to fool. Nature. 2019, 574(7777): 163-166. doi: 10.1038/d41586-019-03013-5
22. Taylor GR. Integrating quantitative and qualitative methods in research. University press of America. 2005, doi: 10.4236/psych.2020.115053
23. McEvoy P, Richards D. A critical realist rationale for using a combination of quantitative and qualitative methods. Journal of Research in Nursing. 2006, 11(1): 66-78. doi: 10.1177/1744987106060192
24. Open Speech and Language Resources. OpenSLR (Version 1) [Dev-clean]. Available online: http://www.openslr.org/12/ (accessed on 7 September 2023).
25. Nasr MA, Abd-Elnaby M, El-Fishawy AS, et al. Speaker identification based on normalized pitch frequency and Mel Frequency Cepstral Coefficients. International Journal of Speech Technology. 2018, 21(4): 941-951. doi: 10.1007/s10772-18-9524-7
26. Hossin M, M.N S. A Review on Evaluation Metrics for Data Classification Evaluations. International Journal of Data Mining & Knowledge Management Process. 2015, 5(2): 3-5. doi: 10.5121/ijdkp.2015.5201
27. Han J, Gondro C, Reid K, et al. Heuristic hyperparameter optimization of deep learning models for genomic prediction. De Koning DJ, ed. G3 Genes|Genomes|Genetics. 2021, 11(7). doi: 10.1093/g3journal/jkab032
28. Mamyrbayev O, Mekebayev N, Turdalyuly M, et al. Voice Identification Using Classification Algorithms. Intelligent System and Computing. Published online April 29, 2020. doi: 10.5772/intechopen.88239